



# divine<sup>™</sup> Northern Light Technology

## **divine SEARCH AND CONTENT INTEGRATION: ARCHITECTURE WHITE PAPER**

---

Northern Light is an advanced search service offered by divine Search and Content Integration (dSCI). This service was developed by Northern Light Technology, which was acquired by divine inc. in January 2002. Today, the Northern Light search service powers divine's Search and Content Integration products.

## TABLE OF CONTENTS

---

HISTORY .....	3
BUSINESS FOCUS .....	4
OVERVIEW OF PRODUCTS AND SERVICES .....	6
DATA MODEL .....	8
SERVICE DESCRIPTION .....	10
DATA COLLECTION AND WEB CRAWLING .....	10
QUERY DATABASE .....	11
Automatic Classification .....	11
Indexing .....	12
Query Service .....	13
SEARCHING .....	13
RELEVANCY RANKING .....	15
CUSTOM SEARCH FOLDERS .....	15
NORTHERNLIGHT VS. NLRESEARCH .....	16
SESSION MANAGEMENT .....	16
E-COMMERCE SYSTEM .....	16
Document Delivery .....	16
Account Types .....	17
User Access .....	17
Account Services .....	18
Shopping Cart .....	18
CUSTOMER DATABASE .....	18
SERVICE ARCHITECTURE .....	20
ALERTS .....	20
GEO-SEARCHING .....	21
APPLICATIONS DEVELOPMENT AND HOSTING .....	21
AWARDS AND PUBLIC RECOGNITION .....	22

## HISTORY

---

Information retrieval/CD-ROM production company Dataware Technology founded Northern Light (now divine Search and Content Integration) in late 1995 with three objectives: (i) unify all of the best content in the world into one database, (ii) build the technology that allows searchers to find easily the most relevant (not just the most) information within that database, and (iii) create a set of capabilities to allow the company to build and operate custom information solutions for businesses. This mission was formed from the following observations:

- Given the success of the Internet, there is no longer any technical or distribution barrier to making all digital information available from any desktop computer. This is not to say that the goal is to replace books or other means of communication. Rather, the intent is to enable such access when needed or wanted.
- As good as the World Wide Web appears to be, it represents only a fraction of all "digitally available data" and much of this data (for example, consumer and professional journals, newspapers, books, economic data, and reference works) are not available on the Web. The presence of this non-Web data was obvious to Dataware, the leading producer of professional-quality CD information at the time.
- First generation Internet search engines have done a good job and advanced the state of the art but they have ignored a considerable body of information retrieval technology that has existed for a number of years, and they do not seem focused on bringing the best search technology (as opposed to the most Web documents) to their products. In particular, the key problem for search engine information retrieval is that of producing a precise set of relevant documents -- fewer good documents, not more useless ones.

To accomplish its mission, Northern Light set out to meet the following goals:

- Build a continually growing database of all the Web information it could find and all the non-Web information it could license, initially focusing on full text and covering all subject areas (academic, trade, popular, medical, business, etc.) and sources (journals, books, newspapers, real-time newswires, etc.). divine's main Northern Light database today supports full text search of every word of 350 million unique Web documents and 30 million unique non-Web documents from more than 7,100 sources,
- Make use of the best existing technology and develop new technology for highly scalable and precise searching, and
- Develop highly scalable automated classification and related technologies to use pre- and post-search because even the best query interpretation and relevancy ranking are frequently inadequate to answer an information need expressed in one or a few words against a database of a billion documents or more.

NorthernLight.com formally launched in August, 1997, and was the first Internet search engine to offer access to both Web and non-Web content.

divine™

## BUSINESS FOCUS

---

The initial strategy was to provide a service to site visitors that would generate revenue from advertising and from sales of licensed, non-Web documents known as the "Special Collection."

The product [NorthernLight.com] was warmly received by existing consumers of high-quality, fee-based information: professionals in corporations, educational institutions and governments. As a result, a marketing effort was directed toward the enterprise market through high volume sales to organizations.

Northern Light's unique focus on both Web and non-Web data coupled with its ability to sell non-Web documents made it of interest to organizations that wanted to be directly involved in offering their own data or services. As a result, the initial strategy was combined with a "gateway partnership" component that typically paired Northern Light with other organizations in the offering of co-branded, specialized search Internet-based sites.

divine Search and Content Integration's enterprise strategy now includes a suite of customized search solutions based on the Northern Light technology that create value for a large range of organizations and situations. These solutions include:

- Custom intranet portals,
- Search-based services for extranets,
- Hosting and sale of archival documents for publishers,
- Search-of-site for Web-based services, and
- Full-scale custom information products (e.g., for the U.S. government), etc.

These information management products reflect the company's core competencies in search technology, classification and taxonomy development, and integration of diverse content and federated search. divine Search and Content Integration typically offers these services on an outsourced, ASP basis. However, the Northern Light search technology is also available as licensed software for in-house customer use on the Compaq Alpha/VMS platform and as a hybrid of Northern Light components and some third-party components on a wide range of platforms. A 100% Northern Light solution will soon be available on platforms other than Alpha/VMS (such as UNIX and Microsoft NT).

divine Search and Content Integration's revenue comes from three sources:

1. Sales of documents to individuals and enterprises
2. Custom search solutions (including ASP search fees, custom development, in-house software licenses, etc.)
3. Gateway partnerships

Revenue is both consumer based (document sales) and enterprise-based (document sales and custom search products).

divine Search and Content Integration's competition comes from three different areas:

- Internet search engines such as Google, AltaVista, or Lycos (Northern Light is unique in its ability to offer a comprehensive database of Web and non-Web documents.)
- Traditional online services such as Dialog, Lexis-Nexis and Factiva that have been in the business of selling high quality (and usually high-priced) documents for many years (divine Search and Content Integration can offer lower prices, newer and superior technology and comprehensive application development capabilities.)
- Custom search and portal providers such as Verity, Autonomy, Plumtree and others (divine Search and Content Integration distinguishes itself by its superior search and relevancy ranking and by the range of its capabilities, including working automated classification and not just classification tools, a complete back office accounting system, and search alerts, among other offerings.)

## OVERVIEW OF PRODUCTS AND SERVICES

---

divine Search and Content Integration offers a consumer-focused search service based at <http://www.northernlight.com> and a corporate-focused search service at <http://nlresearch.northernlight.com/>. Both services allow the user to search Northern Light's database of non-Web documents, with NLResearch.com also allowing search of Web documents. Searches can be easily and directly entered on the one-field, initial "simple search" form or in another of the specialized, multi-field search forms tailored to specific sorts of searches.

NorthernLight.com currently has five search forms:

- **Simple Search** — a single search box on Northern Light's home page
- **Power Search** — advanced searching by subject, source, document type and date
- **Business Search** — specialized searching of industry-focused Web and Special Collection content
- **Investext** — access to thousands of investment analyst reports
- **Current News Search** — search of the database of current news from the last two weeks

NLResearch.com has these five, and three additional forms:

- **Market Research** — access to reports produced by more than sixty consulting and market research firms
- **EIU Search** — search of the Economist Intelligence Unit research reports
- **GeoSearch** — links to Web sites based on their geographic location

Searches can be entered in any form – keyword, Boolean, natural language, +/- syntax, etc. After the search has been performed, the user is presented with a "results list" of documents matching the search request, a brief summary and other information about the document, as well as a hot link to access the full document.

If the document selected for access is a Web document, the user is sent directly to that Web site. If the document selected for access is a Special Collection (non-Web) document, the user is first taken to a free abstract of the document and then given the chance to purchase and access the complete document, which is usually hosted at divine Search and Content Integration. Non-Web documents generally have a fee associated with them (typically between one and four dollars but sometimes as high as thousands of dollars). In order to purchase a document, the user must login to an account, as detailed below. The Special Collection consists almost entirely of full text documents, compared with some traditional search services that store, index and search only document citations and/or abstracts.

Results lists are presented in relevancy-ranked order, by default, and are generally accompanied by a set of Custom Search Folders' (CSFs). These folders sub-divide the results into the key subjects or other categories associated with the search. The user can navigate more quickly through the results list by selecting a category or folder of interest, which contains the documents on the results list that belong to that category, as well as a more specific set of categories. CSFs also serve to disambiguate queries, e.g., a search on 'bonds'

could bring back CSFs for municipal bonds, James Bond, chemical bonds, etc. CSFs are formed from classification information that dSCI collects for every document loaded into its database, and are used to help users limit a search to a specific area.

To purchase and display a Special Collection document, a user must create an account or "log in" to an existing one. Individual accounts, with payment by credit card, can be set up online in a few steps. An enterprise account can be established through divine's Northern Light sales and customer support organizations and can use a number of different payment schemes (most commonly, annual subscription or deposit accounts). This type of account allows a company to administer individual accounts, set spending limits and configure other access options, including on-the-fly IP validation.

divine provides many other services using the Northern Light technology, including:

- current news headlines (presented on the "Current News" search form),
- a form to submit URLs for inclusion in the Northern Light database,
- RivalEye' (resource collections around a specific topic, available from a specific micro-site on NorthernLight.com, NLRResearch.com, or developed on a custom basis for an enterprise client and accessed through their internal intranet), and extensive online help.

A feature unique to divine Search and Content Integration, Search Alerts', which allow users to track topics of interest, deserves special mention. A user can set up an account by entering an e-mail address, and can choose to "save" any search as an alert directly from the results list. Once a Search Alert has been set up on a given topic, the user is sent an e-mail whenever the Northern Light database or any databases searched are updated with new information on that topic. The e-mail includes a direct URL link to the new results list, allowing the user to view the new items immediately. Currently, Search Alerts are used by consumers on NorthernLight.com, by enterprise customers on NLRResearch.com and are included in a number of the custom search services offered by divine Search and Content Integration.

Today divine Search and Content Integration's enterprise products and services include:

- 1) Content products such as the enterprise subscriptions to the Special Collection and other premium content sources including Investext financial analyst reports, EIU econometric reports, market research reports, and others. These are typically flat-rate annual licenses for unlimited access at the desktop.
- 2) Various editorial services including customized RivalEye' (a customized competitive intelligence micro-site) and Live Query' (a dynamic search customized to the client's requirements) products. These editorial services are often combined with content licenses.
- 3) Internet, intranet and extranet search services such as Search Toolkit' that provide the Northern Light search technology and custom content licenses for individual Web sites. These services include a basic site search capability with customized crawls, and licensing of the full Northern Light API in conjunction with a license for Internet or extranet access to the Special Collection.

- 4) Customized enterprise information portals, including SinglePoint' service, that provide a single point of access for the company's knowledge workers to all the available information resources. This typically includes relevant Web content and licensed third party content as well as internal content from local servers and data depositories. A single login is provided for all content sources including appropriate rights management and security authorization, as well as the ability to use Northern Light's taxonomy and auto-classification. divine Search and Content Integration currently sells SinglePoint both as a hosted, ASP based solution and as an on-site solution operating behind a company's firewall. In addition, divine Search and Content Integration integrates SinglePoint into the major corporate portal products such as Plumtree, Oracle and Corporate Yahoo.

## DATA MODEL

---

The basic unit of the Northern Light database is a document. Today, all documents are text-based, although eventually documents will include other media. Each document is viewed as a multi-dimensional object that may have one or more values for a number of different fields, attributes, dimensions and/or domains (terms used interchangeably).

One such special field is the display-object – the viewable document itself, generally stored in HTML format (There can also be an "original-object" – the viewable object in the form it was delivered to divine Search and Content Integration.). Other fields, or values for all other fields – title, author, creation date, subject, etc. – are generally referred to as "metadata," -- data about the document. The term "metadata" sometimes applies to the actual values a document may have for each of these fields, as well as to the named fields themselves.

Certain metadata is required for any document, including a display object, title, price, etc. Metadata is also used for Custom Search Folders and other classification-based browsing and searching, and includes key attributes such as subject, type, source, language and region. This metadata is sometimes present in the document itself, and sometimes it is generated by the Northern Light technology's auto-classification capabilities.

Finally, there is additional metadata that may or may not be present for any document. The first two types of metadata are generally both indexed (available for searching on that field) and used within the service in various ways. The third may or may not be indexed, and may or may not be available in the product even if it is indexed. (Sometimes there are not enough data sources with values for that field to make it generally useful, even in a restricted search situation.) This third type includes fields such as company, ticker, author affiliation, or product name. The basic criterion for establishing a metadata field is that it be essentially independent of any other metadata field; a document could have a value for one metadata field, and any other value (or a wide range of values) for other fields. In most, though not all, cases, a document can have multiple values for a single field – e.g., multiple subjects, multiple authors.

The metadata used for Custom Search Folders – subject, type, source, language and region – is treated specially in a number of ways. The possible values for each of these fields have been defined and comprise a taxonomy or set of possible values for that field/domain. These taxonomies are all hierarchical, though they contain many cross-references; a given value may have more than one parent because each taxonomy is actually a directed acyclic graph.

The subject taxonomy contains approximately 16,000 values (referred to as "nodes"), starting at the top level with broad categories such as 'humanities', and sometimes going eight or more levels deep in certain areas to provide very specific subject values such as "works of W.H. Auden" or "robotics".

The "type" field refers to the kind of document – an article (the default and most populated type), a review (with more specific typing of "book review" and others), an editorial, a letter, a report, something for sale, etc.

The "source" field refers to where the document came from, and is either a Web source of some kind (e.g., a Web site, or possibly higher level source node such as "all commercial sites") or a Special Collection source– typically a single journal or book title at the lowest level (e.g., The Economist, or the Boston Herald) or, again, a higher level aggregate (e.g., "journals and magazines", "news articles", etc.).

The "language" field is the predominant language(s) of the document – currently one of English, French, German, Spanish, Italian and unknown (i.e., some other language).

The "region" field specifies a location or locations referenced in the document – a city, country, geographic region, etc.

For type, language and subject, the metadata value(s) attempt to capture what the document is really about (or substantially written in, in the case of language). Multiple values are possible but these are intended to represent true multi-subject documents. In the case of region, the model is slightly different; a document will be tagged with any and all regions that can possibly be identified with a document. The difference between region and these other fields is the way they are used in searching.

This multi-dimensional model is in contrast to a single dimensional model that must rely on repetition within a single domain in order to achieve comprehensive document descriptions. For example, in a single dimensioned design, values like "reviews" or "biographies" could be repeated under all or a very large number of subject values. As the amount of metadata increases, the single (or few) domain model becomes increasingly complex and unwieldy. The Northern Light multi-dimensional model, however, can maintain multiple taxonomies easily and simply and can class and organize documents against them.

## SERVICE DESCRIPTION

---

### DATA COLLECTION AND WEB CRAWLING

Data comes into divine Search and Content Integration's main Northern Light database from two sources: the Web, via the crawler, and from licensed relationships with publishers and other data providers, via custom acquisition and conversion programs. Data for custom search products can also come from internal enterprise sources or third party sources (based on the Web or elsewhere) to which an organization has access but which may not be part of the main database, and which may be subject to various access control rules within the organization. As with the main Northern Light service data, such data comes either through some form of Web crawling or through construction of a custom data acquisition and conversion capability.

Web data is generally retrieved by a crawler that continuously scours the Web for new or updated documents. The crawler collects text (ASCII) and HTML documents from the HTTP (non-secure) accessible Web. Web data is converted to a standard Northern Light format that captures the document itself plus associated metadata, including title, date, etc. The crawler group is responsible for the development and support of the crawler, though the actual daily operation is generally handled by operations.

Non-Web data, whether a part of the licensed Special Collection or third-party content specific to a SinglePoint implementation, is acquired by divine Search and Content Integration through FTP, tape, satellite, serial line feeds or other means. This data is then converted to the standard Northern Light format used for Web data. Since data typically arrives in non-HTML format, part of this conversion involves changing the document text itself (often in tagged ASCII, SGML or other formats) into HTML. The content processing group, with engineering support, is responsible for planning the initial conversion of data, retrieving it (as appropriate) on an ongoing basis, and preparing it for loading into the database. Some data sources are received, converted and loaded automatically and electronically, such as those for Current News. divine Search and Content Integration has, to date, converted over 100 different data formats, including PDF and documents rendered as images; images are processed with programmatic OCR to make the content available for indexing and full text search.

In the case of certain third-party content licenses specifically for one or more SinglePoint implementations (such as market research content from vendors such as Gartner, IDC or Forrester), divine Search and Content Integration keeps the content only as long as it takes to create the necessary indexes. Once the content has been completely indexed, it is discarded, making it impossible for divine Search and Content Integration to re-create the full-text of these content sources.

## QUERY DATABASE

Once data is placed in a standard Northern Light format, it is loaded into a Northern Light database. Today there are actually three primary Northern Light service databases, as well as a much larger number of custom, private databases for various enterprise customers.

1. The main and largest Northern Light database contains all Web and divine Search and Content Integration -licensed non-Web data. This main database is today updated every two to three days. This database currently consists of one or two terabytes of index information (which represents roughly twice that amount of raw document data) and certain summary document information. The main database does not include the actual Web and non-Web documents. Web documents are discarded after indexing, and the non-Web (Special Collection) documents are stored on a separate premium document server (PDS). Data/documents are generally never deleted from this database; it is an historical, incrementally updated database that has grown continuously since product launch.
2. There is a current news database that is automatically updated every two to three minutes, and its size is less than one GB. Data remains in this database for two weeks, after which it is deleted. Currently, this database only contains data from live newsfeeds (AP and Comtex) sent to divine Search and Content Integration through leased lines. All documents in this database are free. Documents in this database are also added to the main database, but are not available for searching until the completion of the (much longer) main database load cycle.
3. A third service database consists solely of Special Collection data. This database is updated daily. The purpose of this database is to provide better currency for the often currency-sensitive documents (e.g., Investext reports) contained in this database.

Database loading consists of several steps:

### ***Automatic Classification***

To deliver automated classification against a huge and heterogeneous data set, the Northern Light technology uses its own classification taxonomies for subject, type (e.g., article, review, FAQ, job listing, etc.) and other document attributes, drawing on existing taxonomies and supplementing them to provide comprehensive coverage for a wide range of users. An automatic system has also been built that uses multiple strategies (e.g., pattern extraction from training documents, co-location analysis, and structural elements) for classifying documents for a given attribute. Both the taxonomies and the automated system have been in production and supporting end users since August, 1997 and are continually being refined to deliver more comprehensive and precise classification and better operational performance. The primary use of classification information (i.e., metadata) at divine Search and Content Integration today is to organize the results (through Custom Search Folders) of a search by appropriate attribute values. This facilitates rapid navigation and some level of automatic query refinement, while allowing more expert users to limit their search initially by some appropriate attribute value. Metadata is also used as one factor (among many) in relevancy ranking.

Subject classification has been designed to classify a document to the one or a small number of subjects from our 16,000+ term subject taxonomy that a document is truly 'about' (vs.

classifying to all subjects that occur in the document). The system can today subject classify approximately 25% of random Web documents (only somewhat less than human editors are able to do) at accuracy rates of from 60%, using the most rigorous metrics, to 90-95% using user/customer appraisals. These coverage and accuracy rates are significantly better for non-Web documents. The system is also capable of classifying millions of documents per day (where average document size is 9K bytes) from the Web and our 7,100+ non-Web sources on a Compaq Alpha computer with six 64-bit Alpha processors and 16Gb of RAM. Classification coverage and accuracy have been realized by continually engineering and extending both known and novel technologies in light of specifically identified problems. Performance levels have been achieved by fully divorcing the logical classification models from their practical implementation and creating data structures appropriate for rapid classification of documents against the large but relatively static taxonomies, patterns and rules that are the basis of the classification process.

Almost all Web data is automatically classified as described above. For Special Collection data, the situation is somewhat more complicated. Much of this data includes classification information from the original publisher or aggregator. In those cases, divine Search and Content Integration's classification group will generally specify a mapping from the publisher's taxonomy to one or more Northern Light taxonomies. This mapping is then used during the initial data conversion to produce classification information for the appropriate metadata fields. Special Collection classification may also in some cases be entered manually, for example where the quantity of data from a particular source is small and/or the subject or other metadata value is easily determined.

Content processing will usually set the document 'type' and 'language' as part of initial data conversion and processing based on knowledge about the particular document being converted. Special Collection data can also be mapped by classification group to certain subjects and/or type in the same way Web sites can. In the case of Special Collection documents not classified manually, automatic classification programs are run to find missing metadata fields among subject, type, source, language and region, or to supply additional metadata.

### ***Indexing***

During indexing, all of the terms in the documents and metadata fields are extracted and indexed into appropriate index structures; searches can be resolved by using these structures and without having to refer to the original documents. This process is exhaustive and comprehensive; all visible terms in the document display object and all appropriate metadata values are indexed. There are no "stop words" or special characters that are not indexed and there is no practical cut-off in terms of document length at which point indexing stops.

All search terms (including those inside quoted phrases) are viewed as nouns and transformed (stemmed) automatically to their common singular form during indexing (and at query time). This allows a search on a singular or plural noun to find occurrences of either. The stemming rules are fairly simple and do not handle most irregular forms, which tend to occur for very common words not generally useful in searching.

All terms are indexed as all lower-case letters. In addition, all terms containing at least one instance of both upper- and lower-case are also indexed in a special case-sensitive index. This allows queries to find all instances of a term regardless of case; query terms are also translated to all lower-case for initial query resolution. This also allows a match in case-sensitivity with a query term to be used as a relevancy factor.

The above rules are English-language dependent. However, given the symmetry with which they are applied at indexing and query time, they generally preserve appropriate search processing for all languages. True international support, which is upcoming, will include language-sensitive stemming and other sorts of processing.

Numeric tokens (or tokens of mixed letters and numbers) are indexed as text. The system is able to store numeric information as numeric fields (and perform appropriate operations on this date such as finding a maximum value for a field), but that capability is not currently used.

Proximity information can be represented in indices in various ways, allowing either very fast access of short phrases, or complete and precise (but slower) access of phrases of unlimited length.

### ***Query Service***

Finished databases are connected to divine Search and Content Integration's Northern Light network by the "Query Listener" (QL). The QL accepts queries from external clients (such as a Web server) and passes them to the "Query Server" (QS). The QS translates the search syntax and other parameters, queries the database indices, and returns the appropriate citation information and metadata. The Query Listener is also responsible for identifying itself by broadcasting, via UDP, a database identifier and load information. Clients use this information to select a listener appropriate to their mission.

## **SEARCHING**

Nearly all search fields on all search forms accept and process searches in the same way. The query interpretation algorithm proceeds as follows:

1. If the query is well-formed Boolean expression, it is rigorously interpreted as such. Boolean expressions can contain AND, OR, NOT, simple terms ("words"), quoted phrases, wildcards and parentheses including sub-expressions, and may contain an unlimited amount of nesting. In addition, a Boolean expression may itself contain any number of fielded sub-expressions that specify a search against a particular metadata field, e.g., (lawsuit or sue) and title: microsoft or netscape

By default, terms are searched against the 'text' field, which includes all full text and all document metadata. This field may also be specified by use of the 'text:' keyword. Search terms may also include one or more trailing or multi-character wildcards (indicated by '\*') or single-character wildcards ('%') as long as there are at least four non-wildcard characters before the first wildcard, e.g., rachm%inof\*

Search fields are available to the user through appropriate fields on search forms or through the "field:" syntax. Not all fields currently operational in the product are used or documented. Some are for internal testing, or not yet in a form for documentation and external use.

Relational operators can be used for date fields, a feature currently undocumented but available internally. Other specialized pseudo-fields, all generally undocumented, determine various special search options, such as "sort: date" (a reverse chronological sort) or "sort: relevancy" (the default).

2. If a well-formed Boolean expression is not found and the query is more than a specified length (currently 12 terms), a statistical query evaluation process is used. This only requires the presence of a single term in the query for a document to appear on the results list, but makes use of all terms or phrases appearing in a document to determine the best documents. This statistical evaluation can also be forced on a query of any length by preceding the query with the pseudo-field "like:", e.g., like: side effects of anti-depressants and sedatives
3. If neither of the above two conditions is met, then a query with any use of the '+' or '-' operators common among Internet search engines will be interpreted according to generally accepted rules. The rules are that any term or quoted phrase immediately preceded by a '+' must be in a document to appear on the results list, and any term or quoted phrase preceded by a '-' cannot be in any document for it to appear on the results list. Other terms in the query are considered desirable but not required.
4. If the query does not meet any of the above criteria, a "fuzzy" search is performed. This does an implicit AND of most content-bearing words (or what are generally non-content-bearing words if those are the only query terms) but uses all terms entered for relevancy ranking purposes. Some limited natural language analysis is also performed on terms, such as recognition of the word "not."

All query terms are presumed to be nouns and are translated, if necessary, to their singular form using fairly simple algorithms. This allows a match against either a singular or plural form, since all document terms are similarly converted to singular form during indexing. Query terms are also translated to lower case in order to be able to match any form of the word in any document; all document terms are similarly converted to lower-case at indexing time. Mixed case terms are searched against a special mixed case index to provide information about case-sensitive matches for relevancy ranking.

## RELEVANCY RANKING

One of the strengths of divine Search and Content Integration's Northern Light technology is its advanced relevancy ranking algorithms. These not only provide a novel approach to ranking but are based on highly optimized index structures and algorithms that allow divine Search and Content Integration to perform significant relevancy ranking operations on a very large database.

Ranking takes into account several different factors, each of which contributes weight to a document's overall relevancy score and to its eventual placement in the results list. A maximum theoretical relevancy score is calculated for every query, and displayed relevancy scores represent a simple transformation to a 1-99% range of the actual document score as compared to the maximum theoretical score. These factors include the following:

- Number of occurrences of matching terms (term frequency factor, or TF).
- Relative frequency of those terms in the entire database (term inverse document frequency, or IDF).
- Implicit phrase recognition
- Location of matching terms and phrases
- Number and authority of external sites linking to this document (applies to Web documents only)
- Date of the document (All other things being equal, more recent documents are considered to be more relevant than older documents.)
- Classification metadata
- Document length
- Presence of any detectable 'spam'

## CUSTOM SEARCH FOLDERS

For any result set containing more than 25 results, divine Search and Content Integration determines a set of Custom Search Folders (CSFs) before returning the results to the user.

To do this, divine Search and Content Integration examines the metadata values of the documents on the results list, uses those values to determine appropriate CSFs, weighs each of the CSFs and then displays the top-weighted CSFs. divine Search and Content Integration will always try to display a "Special Collection Documents" CSF as the first in a series, to review just the Special Collection part of the results list. The second CSF slot is generally reserved on non-news searches to a 'search current news' pseudo-CSF that launches the same search against the current news database.

Weighting of CSFs is determined by a number of rules that contribute different values to the overall weighting of that CSF. Certain rules assign weights based on how many documents are in the CSF, or how many of its documents rank high on the results list. Other CSFs assign values based on how different the CSFs are from other candidate CSFs, or based on the more exact nature of the metadata values involved.

## **NORTHERNLIGHT VS. NLRESEARCH**

In an effort to offer a tailored set of search options to divine Search and Content Integration's different user communities, there are two Northern Light sites: [www.northernlight.com](http://www.northernlight.com) and [nlresearch.northernlight.com](http://nlresearch.northernlight.com). The two sites are similar and offer the same overall functionality. There are several differences between Northern Light and NLResearch (as they are referred to), the most significant of these being the fact that NorthernLight.com searches only the Special Collection and News databases. NLResearch defaults to Special Collection, but allows searching of Web content, News, and other premium content.

## **SESSION MANAGEMENT**

Divine's Northern Light service uses a proprietary "session state" management system to store user state between page requests or other transactions.

Each user is given a cookie with a unique token that contains no outwardly useful information. The user's browser transmits the token in the 'headers' of each request, and the Northern Light software uses the token to retrieve session data.

## **E-COMMERCE SYSTEM**

A complete commerce system that includes support for secure Web transactions (HTTPS) via SSL encryption facilitates the purchase and delivery of documents.

### ***Document Delivery***

The system (frequently referred to as the "Back Office") supports a wide range of user interactions. The most basic function is the display of Special Collection abstracts, including appropriate metadata. For non-enterprise users, the abstract includes "Purchase Document Now" buttons. Users of enterprise accounts who are presumed to be operating under an "all you can eat" purchase plan have buttons which say "Read Document Now."

Special Collection abstracts may also contain links to other documents. This is most commonly seen in Investext documents, which are sold by the page and by the complete report. Abstracts for individual pages, for example, include a table of contents with links to every other page in the report and a link to the full report. Documents in Adobe PDF format include a link to Adobe's Web site in order to download and install a PDF reader.

### ***Account Types***

Purchasing and viewing the full-text of Special Collection documents requires access to a divine Search and Content Integration account of some type. The commerce system currently supports four different account types: individual, promotional, enterprise and IP-validated. The first type can be created "on the fly" either directly or in the course of attempting to purchase a document by providing appropriate credit card information for charging the purchases. Promotional accounts can be created only through a particular URL but do not require a credit card, instead making use of Northern Light "promotional dollars". The latter two types, enterprise accounts, and IP-validated can be set up and configured by Northern Light customer service, bulk-loaded via a database importation procedure, or via special scripts that create secure accounts on the fly for users who access Northern Light through a third-party edition.

Enterprise Accounts (EA) are essentially collections of user accounts, each of which is billed, in aggregate, to a corporation via a purchase order or other methods. The billing system (QuickBooks) is used to manage a variety of pricing plans and options. User accounts can be configured to have monthly limits, or to require the entry of project or other tracking codes with the purchase of each document. Limits can also be set on a per-source basis, allowing enterprise customers to limit the purchase of the most expensive content sources.

IP-validated accounts are ordinary EAs that use the client's IP address to bill particular user accounts without requiring entry of a username and password. Alternately, IP-validation can permit a client to create a user account under the EA.

### ***User Access***

Unless using an IP-validated account, users must enter a username and password to purchase Special Collection documents.

The commerce system regularly offers the user the option to connect to a secure server via the HTTPS protocol. This connection does not alter the functionality available to the user, but it does ensure that the transmission of credit card numbers, transaction histories, or even usernames and passwords, are not easily intercepted.

Once users have logged on, their session record is updated so that they can stay logged on for 60 minutes.

Users who are logged on and purchase a document from the abstract are immediately brought to the full text. In all cases, the metadata shown on the abstract is repeated above the full text. Several other features, including a "Format for Print" button, are available on the document screen, and the user may also return to their results list.

### ***Account Services***

All account holders can log in to the "Account Services" system and receive a summary of their account, including the last charges billed to their credit card. At this point they can also see their previous transactions, electronically refund documents that did not meet their needs, and update account information such as password and credit card number.

Enterprise Account administrators have access to additional functionality which allows them to administer, in similar fashion, the "users" who are attached to their corporate-level account. Administrators can create new users, limit their purchases on a monthly basis (or per source), and require the entry of project or other tracking information for each document purchased. (The latter is done on a per-purchase basis, and a set of standard codes can be supported for any single enterprise).

### ***Shopping Cart***

Results lists on NLRsearch include an interface to add or remove Special Collection documents to a virtual "Shopping Cart." Additional buttons allow the user to view their cart through the commerce system. The Shopping Cart interface allows review of abstracts, removal of items from the cart, and easy purchase and delivery of all items in the cart.

The back office system also supports:

- Automated refunds
- Multiple subscription plans (none is currently in operation on the main service)
- Promotional (vs. real) dollars (keeping track so as to allow refunds of documents bought with promotional dollars to only be refunded with promotional dollars)
- Full on-line transaction history

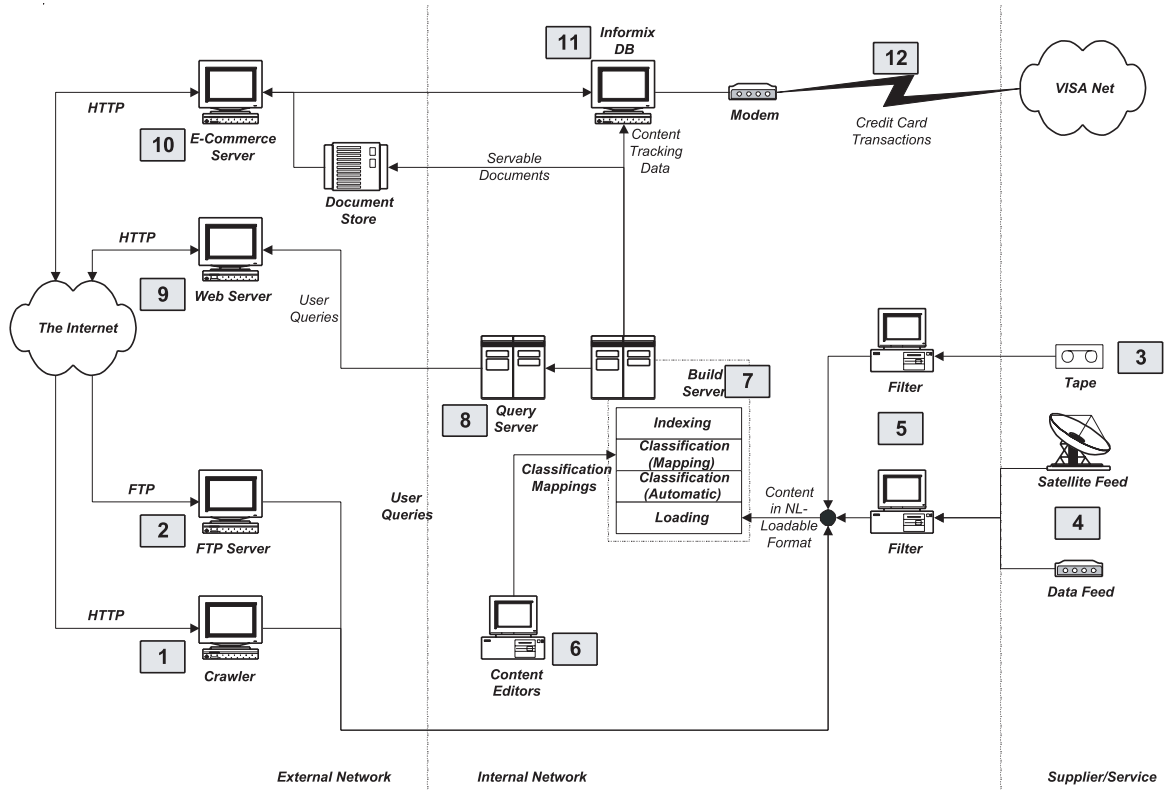
### **CUSTOMER DATABASE**

All customer data, including usernames, passwords and credit cards, as well as commerce transactions (document purchases) are stored in an Informix database. This database is accessed via a proprietary protocol that hides database schema particulars from the commerce system software.

This protocol is also a key part of divine Search and Content Integration's security system. It does not include any transactions which could be used to obtain credit card numbers. This prevents the accidental or deliberate viewing of credit card numbers through the commerce system interface.

Direct access to the database server is extremely restricted, and in addition, all credit card numbers are encrypted prior to storage.

Figure 1. Northern Light Service Architecture



## SERVICE ARCHITECTURE

Please refer to Figure 1 on the previous page.

1. **Crawler.** The crawler continuously scours the Web for new or updated pages. The full text of those pages is passed to the loader.
2. **FTP Site.** Content suppliers deliver Special Collection data in many different formats via Northern Light's FTP site.
3. **Tape.** Content suppliers also deliver Special Collection data on tape, again in any format.
4. **Satellite/Data Feed.** Content suppliers also deliver Special Collection data, particularly news, continuously via direct channels.

In addition, documents on a results list from the same Web site or Special Collection source (e.g., a specific journal title) are generally 'in-line clustered' in the results list, meaning that only the first or first few documents are shown (depending on how good the 2nd and 3rd documents are perceived to be) and a link represented as a Custom Search Folder is present to access the remainder of the document from that Web site or source. This ensures that a single content source does not completely dominate the initial results while still providing immediate access to multiple documents from the same source.

## ALERTS

divine Search and Content Integration offers users and enterprise customers the ability to save any search and have it run automatically whenever the database referenced by the search is updated. At that time, an e-mail is sent to the registered owner of the alert if and only if there are new documents in the database that meet the search criteria; the e-mail message contains a link to just these new results. The system further keeps track of when a user has actually accessed these new results so that, if a user receives a string of alert e-mails before being able to see any of them, accessing any one of them will provide all the new results since the user's last access; it is unnecessary to cycle through the alert messages one at a time in order to see all new results.

## **GEO-SEARCHING**

divine Search and Content Integration geo-codes all Web documents that it loads, translating any U.S. location information (an address, telephone number, city, etc.) into a latitude-longitude point representation. Users choosing to perform a geo-enabled search can supply corresponding location requirements on the search (e.g., within 5 miles of a specific location), and the system will return only results that contain matching address information.

Applications development and hosting

divine Search and Content Integration offers complete services for producing, or letting customers produce, custom search applications to be run either in ASP mode, in-house at a customer site, or some combination of the two. These can include documented APIs for searching, customized results lists, alerts and other capabilities (usually through XML interfaces). divine Search and Content Integration also has a dedicated applications group, efficient tools for rapid development of user interfaces and, behind it all, a 7x24 secure operations facility.

## **APPLICATIONS DEVELOPMENT AND HOSTING**

divine Search and Content Integration offers complete services for producing, or letting customers produce, custom search applications to be run either in ASP mode, in-house at a customer site, or some combination of the two. These can include documented APIs for searching, customized results lists, alerts and other capabilities (usually through XML interfaces). divine Search and Content Integration also has a dedicated applications group, efficient tools for rapid development of user interfaces and, behind it all, a 7x24 secure operations facility.

## AWARDS AND PUBLIC RECOGNITION

---

- "Top 100" eContent magazine, December 2001
- "Best of the Web" US News and World Report, October 2001
- "Top 100 Companies That Matter" KMWorld magazine September 2001
- "Editors' Choice" PC Magazine, November 2000
- "Best of the Web", Forbes magazine, September 2000
- "Web Business Award For Online Excellence" CIO magazine, July 2000
- "Best Online Business/Professional Service", Software & Information Industry Association, March 2000
- "Best Online Research Product", Software & Information Industry Association, March 2000
- "Best Online Information Service", Software & Information Industry Association, March 2000
- "Editors' Choice" PC Magazine, September 1999